THE STANDARD ECONOMIC PROCESSING SYSTEM: A GENERALIZED INTEGRATED SYSTEM FOR SURVEY PROCESSING*

Shirin A. Ahmed and Deborah L. Tasky, U.S. Bureau of the Census
Deborah L. Tasky, U.S. Bureau of the Census, Washington, D.C.

Key Words: Economic surveys, Standards, Generalized systems, Collection, Post-collection

INTRODUCTION

The economic area at the U.S. Bureau of the Census conducts 110 current surveys in the areas of retail, wholesale, service industries, manufacturing, and construction. Prior to 1995, subject areas directed the development of systems to accommodate specific program needs. Over time, this resulted in the economic area having 16 different processing systems, and variations of each. A review of these systems showed[1] the following:

- Many systems performed similar functions. In fact, in many instances, differences among systems could not be explained.
- Separate resources maintained and managed each system. This meant multiple groups were solving similar processing problems. Additionally, processing problems distracted subject matter resources away from critical program needs.
- Subject matter areas with more resources had better systems -- that is, systems with more functionality. Enhancements were program-specific.

In May 1995, the economic area dedicated a team to build a common survey processing system. The team comprised survey statisticians, programmers, and mathematical statisticians. The system they developed is known as the Standard Economic Processing System, or StEPS. In four years the team completed the basic StEPS system. During 1999, 50 annual surveys used StEPS to collect and process data for the 1998 statistical year. Note, three of the surveys had served as StEPS pilots in the prior year.

In the next three years, the remaining surveys move to StEPS. At that time, approximately 400 analysts at headquarters in Washington, D.C., and 80 clerks in Jeffersonville, Indiana (site of the Bureau's processing center) will use StEPS.

ABOUT ECONOMIC SURVEYS

The economic area's surveys that will migrate to StEPS represent annual, quarterly, and monthly programs. Many surveys have as their frame either the business register (i.e., the Standard Statistical Establishment List) or derivative files from the business register. Most surveys are establishment or company based. Exceptions are the construction monthly surveys, which are project based, and whose frame is outside the business register.

The first phase of StEPS covers establishment and company-based programs. These surveys are primarily mailout/mailback programs, which follow-up with respondents using a variety of techniques, including Computer Assisted Survey Information (CASIC) technologies. Once data are captured, computer systems process it through post-collection activities for data cleaning, estimation, analysis, and dissemination.

WHAT IS StEPS?

It follows that StEPS is a processing system that contains the following:

- *Standard data set structures* to support all aspects of survey processing.
- Integrated modules that perform —
  - *administrative functions* that let users modify StEPS to meet survey requirements.
  - *post-collection processes* such as editing, imputation, data review and correction, data query, estimation, analysis tools, disclosure, and variance estimation.
  - *support functions for collection technologies* that include mailout, check-in, data capture, and follow-up.
  - *linkages to outside systems* such as the business register.

*This paper reports the results of analysis by Census Bureau staff. It underwent a more limited review than official Census Bureau Publications. This report is to inform interested parties and encourage discussion.

Regarding the linkages to outside systems, activities outside StEPS cover frame development, sample selection, actual CASIC technologies, and data dissemination.

SAS® AS THE FOUNDATION

The standard data set structures and integrated modules in StEPS are developed using the SAS language. Choosing SAS as the foundation for StEPS came on the heels of two significant events. First, a pilot system built for the Farm, Ranch, and Irrigation Survey (FRIS)[2] showed promise in its application of SAS as a development tool. Prior to then, many only saw the benefits of SAS software for statistical analysis. Second, the Bureau purchased a site license for SAS.[3] The site license gave Bureau staff open access, training, and in-house support to multiple products provided by the SAS system.
Since SAS runs on many platforms, SAS fits in with Bureau objectives to move to open systems. For the economic surveys, StEPS is configured for the Unix operating system (on DEC Alpha machines). Users access StEPS via a graphical (X-windows) communications emulation package loaded onto their microcomputers.

The decision to use SAS had ancillary benefits. As a fourth generation language, SAS became easy-to-learn and easy-to-use compared to traditional languages such as Fortran, COBOL, or C. This reduced development time for programmers. Additionally, it paved the way for survey statisticians and mathematical statisticians to learn a common language, thus fostering communication within the economic area. Finally, SAS came with self-contained products that the team exploited. For example, the team took advantage of the following from the SAS system: use of pull-down menus; organization of user-help text; data query capabilities; access to SAS's data analysis package, called SAS/INSIGHT®, and access to SAS's point-and-click analysis package, called SAS/ASSIST®. These products further reduced development time yet expanded functionality for users.

HOW IS StEPS GENERALIZED?

StEPS accommodates a variety of surveys because of its general design. To achieve a general design, the team instituted standards, adopted a "skinny" record structure for information storage, and used parameters as the driving force for fitting surveys on StEPS. Each of these is discussed below.

Standards

Standards fortify the StEPS'general design. They are building blocks for information storage, shared input and output files among modules, graphical user interfaces, and code development. With standards, work by development staff proceeded independently, yet downstream component parts fit together. Thus, standards laid the foundation for rapid development.

Instituting standards was difficult. For the information stored within StEPS, the team worked with the user community to agree on field content, definitions, and values — such as survey name, data flags, data versions, and status codes. The user community consisted of survey statisticians in the subject matter areas, often referred to as survey analysts; mathematical statisticians responsible for imputation, estimation, variance estimation, and methodology issues; and processors, who manage the work flow and translations of survey requirements into StEPS. Representatives from the user community served as "advisory consultants" to the team. Resulting "decision documents" reflected formal issuance of information standards.

For generalized processes, standard input and output file formats exist in StEPS as a means of communicating with the modules. The team's crusade to standardize output from data capture — regardless of technology — impacted many parts of the organization. It required changes in Key Entry III software used for heads-down keying, Computer Assisted Telephone Interviewing (CATI) software, and laser check-in equipment. Standardizing output from these external systems into a StEPS *standard data output* (or SDO) format allowed for *one* batch update program to be written for StEPS. Any update to StEPS from a system external to StEPS is easily handled as long as files are in SDO format. For example, obtaining administrative information from the business register means simply extracting the information, putting it into SDO format, and executing the batch update program.

For user interaction with modules, standards support common function keys, pull-down menus, screen layouts, and field displays for colors, highlighting, and notation. Lastly, in addition to these standards, developers follow guidelines in structuring program code.

Skinny Record Format

The SAS software has its own defined data set structures. StEPS stores information in SAS data set structures. In setting up how to organize these data set structures, the team decided on a "skinny" record format. Skinny means storing survey items and variables as records in a data set. As an example, if a survey collects 100 items, a given case (or ID) has 100 records on the "item data set." A separate record exists for each item.

A skinny design accommodates variations among surveys. Each survey has different items stored as records. Within a survey, if items change from one statistical period to the next, StEPS easily handles the change. The mechanism for this is the Item Data Dictionary. StEPS creates its

item data set directly from the Item Data Dictionary, which users access and update to reflect new survey requirements or changes.

Parameters

That brings us to the next area that makes StEPS general. Parameters drive the StEPS system. They allow a general system to be customized for a particular survey. In StEPS, parameters are thought of differently than those used in traditional survey processing. In StEPS, we refer to the parameters as *survey specifications*. We have different types of survey specifications.

Survey specifications that define survey content are *dictionaries* in StEPS. Two major dictionaries exist. The first defines the items (from the questionnaire) and the derived items. It is called the Item Data Dictionary. Noted earlier, this dictionary is the vehicle for flexibility in content variation.

The second dictionary defines control information -- and is called the Control File Dictionary. Control information is characteristic information about survey cases -- such as name and address -- which is standardized across surveys on StEPS. Also, there exists survey specific control information that varies among surveys.

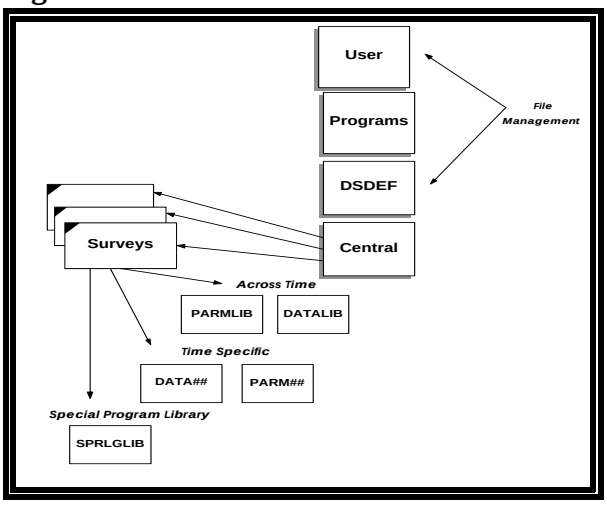
Survey specifications that define business rules for a particular survey are referred to as *definitions* in StEPS . Definitions are used in nearly all modules in StEPS. For example, users define rules for editing data, choices for imputation methods, "where clauses" to select appropriate cases to mail, and table specifications for estimation.

For many operations, the definitions feed into a StEPS code-generator to produce program-specific executable code. For example, in general imputation, the user selects the "ratio method" and indicates the item to be imputed, the numerator and the denominator of the ratio, and the category for making the imputation base (such as industry average). StEPS generates the corresponding code that identifies the items to impute, creates the imputation base, imputes the appropriate item, and sets the appropriate flags. Another example is the edit module. When users enter "required item" tests, StEPS captures the entries and generates code accordingly.

STANDARD DATA SET STRUCTURES

As noted earlier, information is stored in SAS data set structures. The organization of these structures is depicted at a high level in Figure 1, Overview of StEPS Data Set Structures.

Figure 1: Overview of StEPS Data Set Structures



File Management Directories

From Figure 1, several directories house information for *file management*. The *user* directory contains default information reflecting the last choices a user made regarding survey, statistical periods, and printers. Most users of StEPS work primarily on one survey and use StEPS for review of specific statistical periods -- for example, analyzing the 1998 data from the Annual Retail Trade Survey. The *programs* directory contains production code to run a survey in StEPS. Lastly, the *DSDEF* directory stores data set definitions for all data set structures used by StEPS. As surveys migrate to StEPS, this directory contains the shells which are initialized then populated with information from legacy systems

Central Directory

A central directory stores information related to all surveys operating on StEPS. One data set contains a record of every survey that points to that survey's top directory. Other data sets contain available printers and valid codes for economic area-wide control file information common to all surveys.

Surveys Directories

For every survey in StEPS and catalogued in the *central* directory, a *family* of directories exist. The top *surveys* directory houses critical data sets. There is a data set that stores valid statistical periods for the survey. If 20 years of data are stored for an annual survey, there are 20 records on this file. Each record contains the physical location of the other libraries needed for the survey. Storing the survey's physical location in this manner keeps the StEPS code general, since the code never needs to know the location of information; it picks up the

location when it reads this record. Another data set contains users and their privileges for changing parameters, running programs, or updating survey data. A third data set contains the questionnaire numbers (or forms) used by the survey. This data set is used in collection activities. A fourth data set contains the valid control file fields for the survey-wide standard information -- again, like the status codes, State code, and so on. Lastly, a data set keeps a processing log of runs for the survey.

Under the *surveys* directory are groups of directories that support the processing for the survey. The first group contains survey information that *crosses time*. In other words, this information does not get replaced every time the survey turns over to the next statistical period. Data libraries -- referred to as DATALIB in Figure 1 -- contain data sets for master control information. It has a record for every possible respondent in the survey along with control file information.

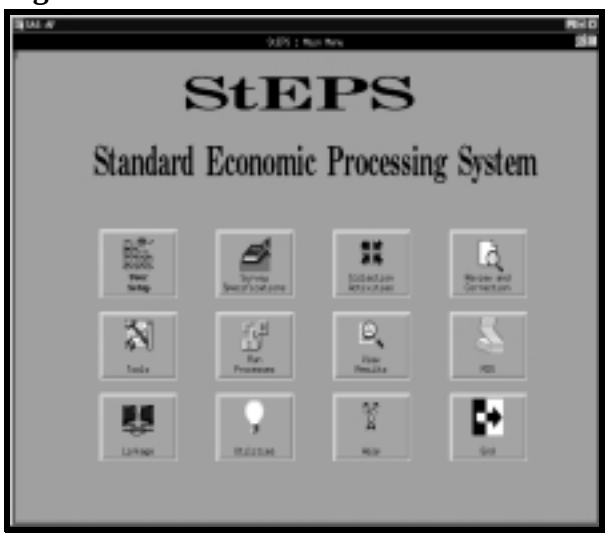
Also, stored are notes about specific cases from a variety of sources, audits, and special mailing group arrangements. Regarding the special mailing group arrangements, StEPS accommodates mailing multiple forms to one location. Lastly DATALIB contains a history of every collection action a case underwent -- for example, mailed, checked-in, and keyed. In the same vein the parameter libraries -- referred to as PARMLIB in Figure 1 -- contain data sets of the parameters in StEPS that do not change each statistical period. Stored in PARMLIB are the survey specifications discussed earlier.

The next group of data sets under the survey directory vary with time or are *time specific*. DATA## contains control file information varying by statistical period. Also stored are item data sets, one for each statistical period of data available for the survey. DATA## contains the survey's edit rejects and any fat record data sets. Fat record data sets are used as input to many processes in StEPS. Because StEPS uses a "skinny" design in data storage, to get efficiency in processing, a fat record structure -- which contains pertinent information for a case ID -- is created as input into processes such as editing and imputation. PARM ## contains more traditional types of parameters used for editing, such as upper and lower bounds for a ratio test.

The last directory under *surveys* is titled SPRLGLIB, which houses survey-specific programs. Found here are customized programs for survey functions not generalized in StEPS. For example, many of the analytical listings are custom coded and stored here. Also stored are code-generated programs for editing and imputation and scripts for survey-specific batch runs. Scripts indicate the order of tasks for a submitted batch run.

MODULES

Figure 2: StEPS Main Menu



Like many systems, StEPS provides an interface to interact with both data set structures and modules. The top level menu is shown in Figure 2, StEPS Main Menu. Menu selections walk users through various modules so knowledge of data set structure and location is not necessary. While an exhaustive review of each module cannot be discussed here, some key points about the modules are discussed below.

Modules to Administer StEPS

Since parameters drive StEPS, these two key modules allow users to specify them: *User Setup* and *Survey Specifications*. In these modules users populate or change parameter files.

Modules to Support Collection Activities

The *Collection* selection lets users generate files for mailing activities, interactively check-in receipts (by mail, fax, or other), and perform batch update operations. As mentioned earlier, batch update operations apply data in SDO format to StEPS data structures. Note, most surveys in StEPS use this program to apply data outputted from Key Entry III (heads down keying) software.

Simple label imprinting is done within StEPS. Use of the new DocuPrint technology -- which imprints both the questionnaire and the variable data -- requires customized information from StEPS for each survey. StEPS provides information for the standard label; the remaining information is obtained by having subject-specific programmers access data set structures directly.

One ancillary collection technology from StEPS is an interactive check-in module. Receipts received in an area away from the laser check-in facility -- such as those received by fax or telephone -- are checked in

interactively.

Modules for Post-collection Activities

The remaining modules relate to post-collection activities. When doing the actual surveys, these are the modules used most of the time, with the primary one being the *Review and Correction* module. The *Review and Correction* module lets users review in various ways the data at a micro level and correct the data. Some options to view data are ID level by item, item by ID, or historic data. Within *Review and Correction*, and through many StEPS modules, users can define a selection set of cases to access by indicating a simple "where clause." For example, users may want to view all cases failing a specific edit test or classified in a specific industry. Also accessible in *Review and Correction* is control file information and capabilities to change coverage (adds, deletes, ghosts, and mergers).

The *Tools* module provides access to SAS products such as SAS ASSIST (user-specified query and tab), and SAS/INSIGHT (for graphical data review). Within *Tools*, the *data query* allows analysts to run "canned" queries. Additionally, through *Tools*, survey analysts can create their own data set to download for further analysis on other desktop software or SAS products. It is this type of feature that makes StEPS enabled to handle "what if" scenarios in survey processing; it gives users the capability to mine the data for anomalies that fall outside normal editing and verification techniques.

Run Processes lets users submit programs that do the following in StEPS: deriving data, editing, imputation, estimation, and, at some future point, mass corrections. When submitted in batch mode, users review results of these processes in the *View Results* module. Some discussion of these modules within *Run Processes* is warranted here.

Deriving data calculates derived items, as specified in the Item Data Dictionary. Editing in StEPS identifies problem cases; the module does not change the data. When selected, the edit module executes the following categories of edits: Required (item), Range, List Directed, Balance, and Survey Rule. The edit tests are entered by users via the *Survey Specifications* module. Most of these tests are self-explanatory. The survey rule test, however, is designed for free form edits that evaluate data relationships. Users write tests in SAS code. For all categories of tests, users specify the "events" where the test occurs. For example, an event is when a user indicates an edit test for a specific case in *Review and Correction*. Users run edits in batch mode or interactively. Running edits in batch requires special privileges.

Imputation is the module that changes data. It has two sub-components. *Simple imputation* changes items, but treats changes as reported. Surveys execute *simple imputation* to prepare items for editing. StEPS designates two types of simple imputation. The "balance complex" executes one of 10 possible conditions for ensuring detailed items agree with the corresponding total item. Users choose the complex -- that is the details, the totals, and the method of replacement -- in *Survey Specifications*. A "free form" options lets users write SAS code to list error conditions and corresponding actions.

The second sub-component is general imputation. It changes data and flags it as imputed. General imputation covers both item replacement or delinquent (form) replacement. As users specify edits, they mark "events" for general imputation, also known as "g-events." Users, generally mathematical statisticians, indicate the method used for imputation -- StEPS offers 21 methods -- and the condition. For indicated methods, an imputation base (i.e., warm deck parameters) is created when the user executes general imputation. At this point in time, general imputation is run in batch mode only.

For estimation, mathematical statisticians specify for "tables" the given "BY variables" and "analysis variables" to produce specific cell definitions. A table is comprised of interior cells for the actual BY variables and margin cells for summary information. This type of specification information is stored on Estimation Specifications Files. All calculations are stored on the Estimation Results File. This file is both an input and output file to variance estimation. StEPS provides generalized modules to handle variances for Poisson sampling, Tillé sampling, or random groups. For samples requiring pseudo-replication methods, the Bureau's VPLX software is used.

Linkages to Outside Systems

At some future point StEPS will link directly with other systems within the economic area. For example, the time series files for the indicator programs or the corporate metadata repository.

MIGRATING SURVEYS TO StEPS

The strategy for migrating surveys to StEPS is to phase in annual and non-indicator surveys followed by indicator surveys. For the annuals and non-indicator surveys, approximately 50 surveys remain for migration. Plans call for migrating 30 surveys in calendar year 2000; followed by the remainder in 2001. Economic indicator programs will follow in 2002.

To facilitate the migration, in addition to the StEPS team, two other groups within the economic area perform important roles. There is a processing group who works with survey statisticians in the program areas to translate requirements into StEPS. For example, they train and help program areas set up item dictionaries, define edit tests, and understand imputation choices. They develop the schedules for the migration effort, assign responsibilities, and track progress. This group is instrumental in specifying how to convert historic data and information from legacy systems to StEPS. They work closely with the last group involved in StEPS -- the subject-specific programmers. Under StEPS, the subject-specific programmers set up the StEPS directories and file structures for each survey. They develop programs that migrate historic data and information from legacy systems to StEPS. Lastly, they work with the processing group to develop any customized programming. An example of customized programming would be analytical worksheets, which tend to be survey specific and difficult to generalize.

FUTURE ENHANCEMENTS

In tandem with migrating surveys, the economic area is pursuing two paths for enhancing the basic StEPS systems. The first is with the users. As they understand and work with the system, users identify improvements in the functionality. The processing group, mentioned above, tracks the user enhancements and researches their viability. The viable ones are presented to a StEPS User Review Board for priority and resource allocation. The User Review Board is comprised of senior program managers. Generally, these enhancements deal with specific changes within modules.

A second path for enhancing StEPS is adding modules so that StEPS becomes an all-encompassing processing system. Over the next three years, development activities will focus on new modules that support electronic reporting systems, perform interactive data collection for clerical staff at Jeffersonville, Indiana, provide expanded macro analytic tools, and prepare data for dissemination.

ENDNOTES

[1] In managing scarce resources, in 1994 the economic area analyzed processing requirements for each program area. The goal was to determine critical "must" activities over the next three years. In the process of doing this, management gained an understanding of the duplication among systems.

[2] Note, in 1996, when the Census of Agriculture transferred from the Bureau of the Census to the National Agricultural Statistical Sevices, FRIS went too.

[3] The final site license for SAS® became effective October 1995 at the Bureau of the Census.

REFERENCES

Ahmed, Shirin A., "Background Paper About the Teams in the Economic Directorate," Census Bureau Internal Document, September 1, 1995.

Chew, Deborah A., "StEPS Glossary," Census Bureau Internal Documentation, January 13, 1999.

Chew, Deborah A., "StEPS Implementation Checklist," Census Bureau Internal Documentation, September 9, 1998.

Luery, Donald M., "General Imputation for Balance Complexes," Census Bureau Internal Documentation, February 18, 1999.

Luery, Donald M., "StEPS General Item Imputation Methods," Census Bureau Internal Documentation, October 20, 1998.

Monahan, James L., "The FRIS Processing System," Census Bureau Internal Documentation, November 20, 1994.

Sigman, Richard S., "How Should We Proceed to Develop Generalized Software for Survey Processing Operations Such as Editing, Imputation, Estimation, etc.?, Proceedings from the Meeting of the Census Advisory Committee of Professional Associations, May 1 - 2, 1997.

Tasky, Deborah L., "Basic Functions within StEPS," Census Bureau Internal Document, January 14, 1999.

Tasky, Deborah L., "StEPS General Imputation Development Activities," Census Bureau Internal Document, July 14, 1998.

Tasky, Deborah L., "Brief Description of StEPS Files, " Census Bureau Internal Document, August 8, 1997.

MORE INFO?

Contact authors via E-mail: *sahmed@census.gov* or *Deborah.Lee.Tasky@ccMail.Census.GOV*